Computer Arabization

Marwa F. Mohamed

Agenda

- Introduction
- Course Objectives
- Course Content
- Course Evaluation
- Projects
- NLP

Arabization

Means the transfer of the foreign word and its meaning into Arabic according to rules and bases guaranteeing its clarity and eloquence.

Computer Arabization

Defined as the conversion of well-tested and well-proven computer software to enable Arabic-speaking users to utilize it effectively and efficiently.

Course Objectives

- The course gives the students a **good understanding** of the **Arabic language characteristics and problems** with programming and different computer applications.
- It gives the students the ability to think in how to localize the software to be convenient to the Arabic users.

Course Content

- **Overview of Arabic language:**
  - Arabic Language **characteristic**,
  - Arabic language **morphology**
- **Arabic Natural Language Processing (ANLP):**
  - Introduction to NLP
  - Challenges in Arabic Natural Language Processing
  - Machine Learning Implementations in Arabic **Text Classification**
  - Arabic Optical **Character Recognition**
  - ANLP applications ..
- **Arabization research activity (project discussion)**

Course Evaluation

Code No.	Course Title	Hours		Marks			
		Lect.	Lab.	Written Exam	Oral &Lab	Test	Total
CS417	Computer Arabization	**2**	**2**	**70**	**15**	**15**	**100**

Projects

Project

- OCR Application
- Speech recognition Application
- Arabic language Education Application
- Augmented reality Application
- Plagiarism Application
- Sentiment Analysis Applications
- Spell checking Application
- Web Clustering Application

Project Evaluation

Documentation	Definition	2
	Challenges of Arabic with this APP	3
	Related work (at least 3)	3
	future work	2
Implementation		5

NLP

Natural language processing (NLP)

- **NLP's** general task is to make a computer understand written language in the form of words, sentences, or paragraphs.
- NLP is for machines and computers to be able to successfully accomplish tasks concerning natural language.
- Natural language refers to any human language that was developed through natural circumstances and follows a specific syntactic and semantic system.

Hapke et al. (2019) defines NLP as:

- an area of research in computer science and artificial intelligence (AI) concerned with processing natural languages such as English or Mandarin.
- This processing generally involves translating natural language into data (numbers) that a computer can use to learn about the world.
- And this understanding of the world is sometimes used to generate natural language text that reflects that understanding.

Natural Language Generation (NLG)

- **Natural Language Generation (NLG),** a subcategory of **Natural Language Processing (NLP),** is a software process that automatically **transforms structured data** into **human-readable text.**
- Using NLG, businesses can generate thousands of pages of data-driven narratives in minutes using the right data in the right format.
- NLG is a subcategory of **content automation** focused on text automation.

Natural Language Understanding (NLU)

- **Natural language understanding (NLU)** is a subfield of natural language processing **(NLP),** which involves **transforming human language** into a **machine-readable format.**
- NLP covers all areas of communication between humans and computers, from input to processing to response.

NLP and Natural Language Understanding (NLU)

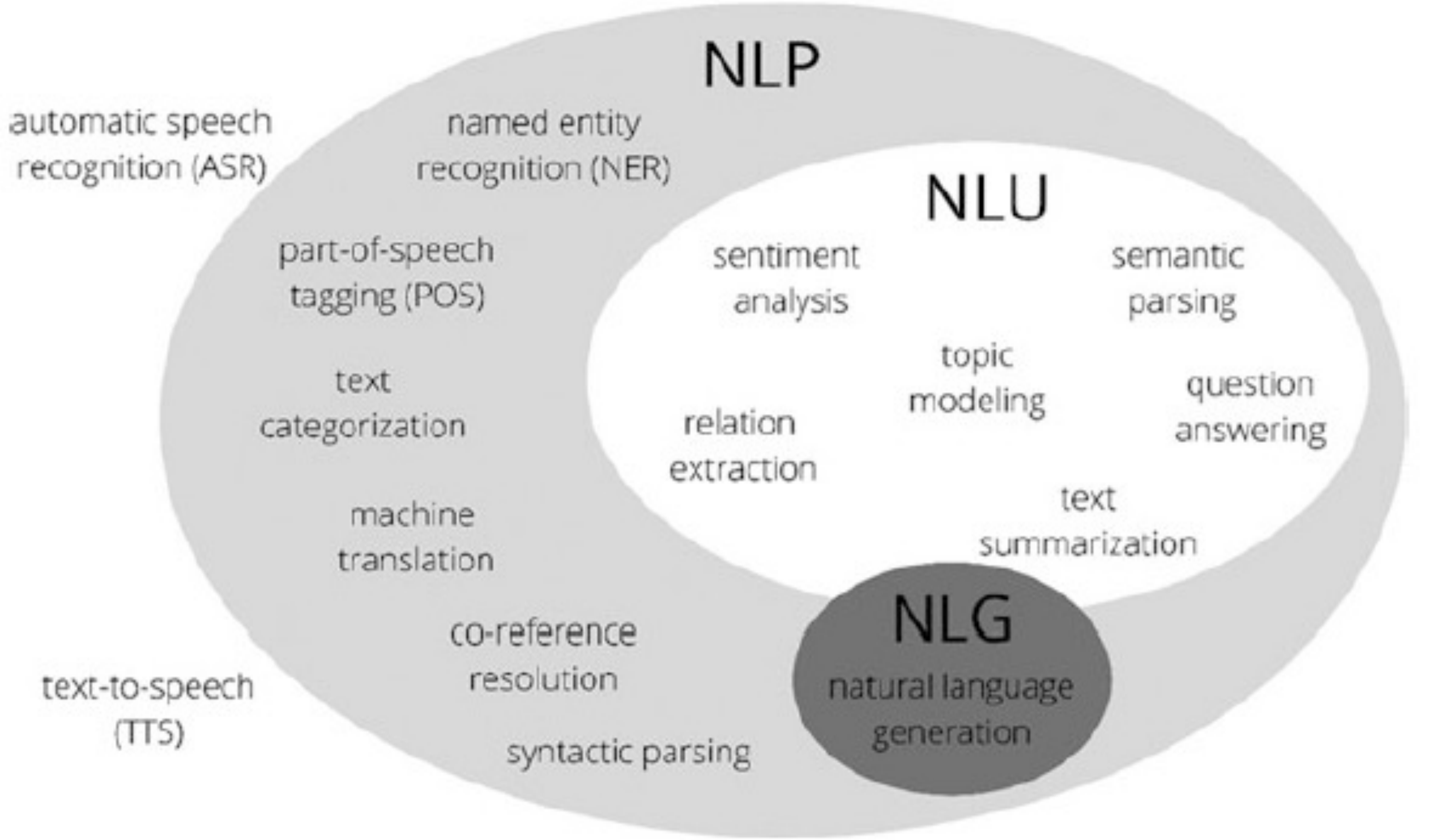


Fig. 1 Terminology relating to NLP, NLU, and NLG. Source: adapted from MacCartney (2014)

NLP Applications

1. Machine Translation
2. Speech Recognition
3. Sentiment Analysis
4. Question Answering
5. Text Classification
6. Character Recognition
7. Spell Checking

- Automatic Summarization

Machine Translation

- Everyone knows what is a manual translation — we translate information from one language into another.
- When the same thing is done by a machine, we deal with “Machine” Translation.
- The idea behind MT is simple — to develop computer algorithms to allow automatically translation without any human intervention.
- The best-known application is probably **Google Translate**.

Speech Recognition

- Is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format.

- **Applications**
  - virtual assistance
  - video games

Sentiment Analysis

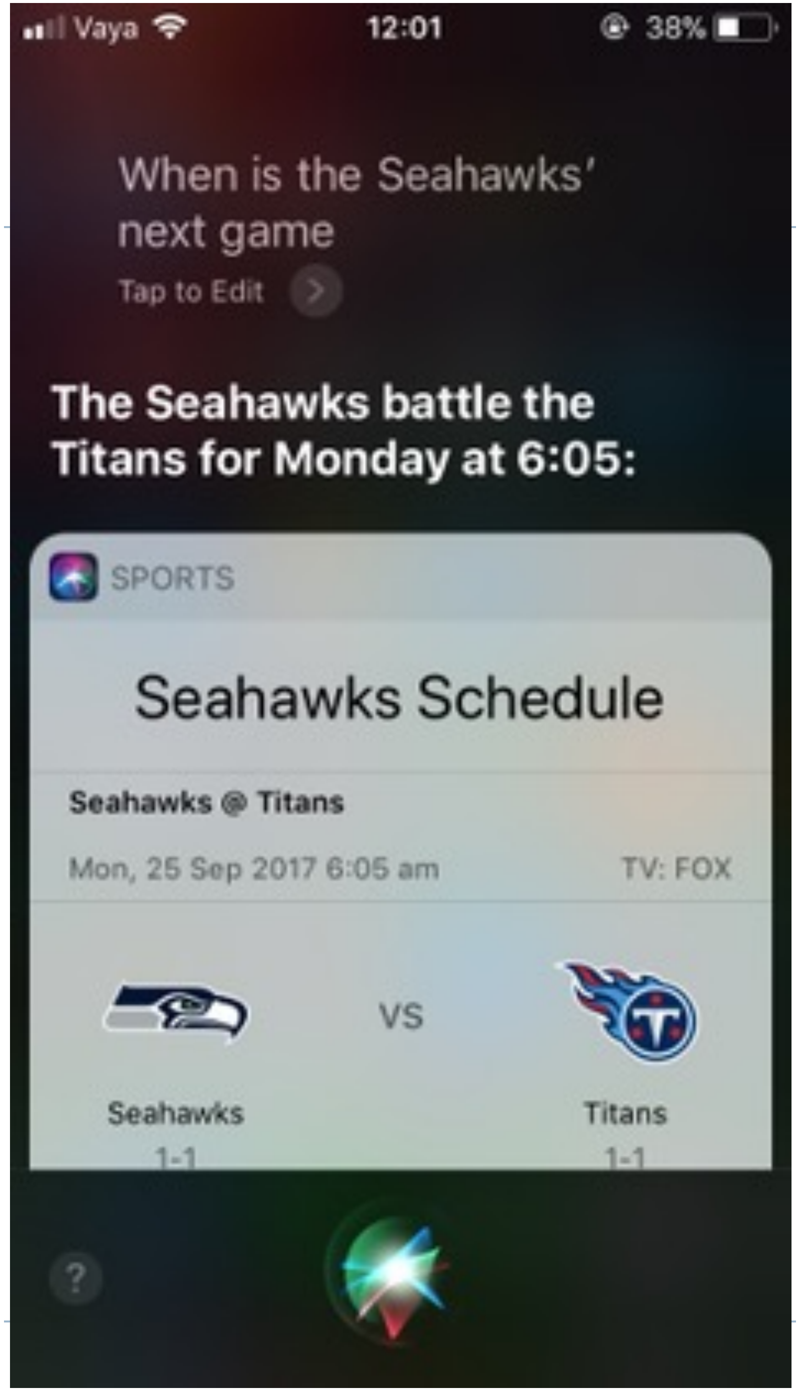
- Also known as opinion mining.
- Is an interesting type of data mining that measures the inclination of people's opinions.
- The task of this analysis is to identify subjective information in the text.

Sentiment Analysis

- For example:
  - this can be a **movie review,** or an emotional state caused by this movie.
  - Why do we need this? Companies use sentiment analysis to keep abreast of their reputation.
  - Sentiment analysis helps to check whether **customers are satisfied with goods or services**.

Question Answering

- is concerned with building systems that automatically answer questions posed by humans in a natural language.
- **Examples** of Question-Answering applications:
  - **Siri,**
  - **OK Google,**
  - **chat boxes**
  - **virtual assistants**.



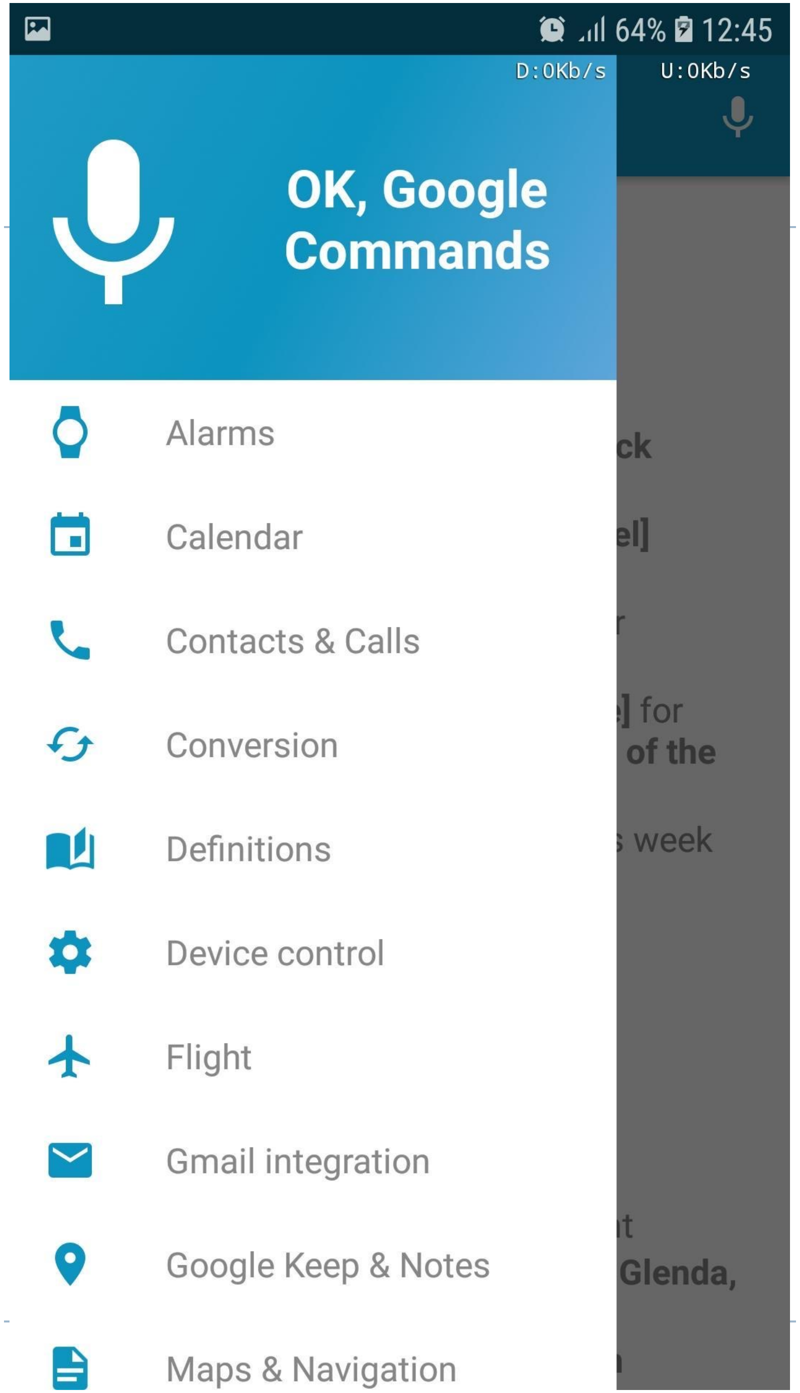
When is the Seahawks' next game

The Seahawks battle the Titans for Monday at 6:05:

Seahawks Schedule

Seahawks @ Titans

Mon, 25 Sep 2017 6:05 am TV: FOX



OK, Google Commands

- Alarms
- Calendar
- Contacts & Calls
- Conversion
- Definitions
- Device control
- Flight
- Gmail integration
- Google Keep & Notes
- Maps & Navigation

Text Classification

- Is the task of automatic sorting of a set of documents or text scripts into categories from **a predefined set**.
- **Examples**
  - A **spam filter**, has emails as its input and classifies them into one of two categories **{spam, non-spam}.**
  - **News stories** are typically organized by topics such as **politics, sports, and economy.**

Optical character recognition

- Is the process of analyzing and identifying texts in scanned images or documents.
- **Examples :**
  - invoice character recognition
  - check character recognition

Spell Checking

- Is a software tool that **identifies and corrects any spelling mistakes in a text.**
- Most text editors let users check if their text contains spelling mistakes.
- **Examples :**
  - **Grammarly app**.

Automatic Summarization

- **Automatic summarization** is the process by a which computer program creates a shortened version of text.
- The product of the process contains the most important points from the original text.
- Search engines such as Google use automatic summarization to produce key phrase extractions in search results.

Match the following technologies with their applications.

Machine Translation C	A. video games
Spell Checking D	B. Siri
Text Classification E	C. Google Translate
Question Answering B	D. Grammarly App
Speech Recognition A	E. Spam Filter

NLP Techniques

1. Text Preparation and Pre-processing
2. Language Detection
3. Tokenisation
4. Lowercasing and Removal of Punctuation
5. Expand Contractions
6. Removal of Stop Words
7. Removal of URLs, HTML Tags, and Emotions/Emojis
8. Correction of Spelling
9. Part of Speech Tagging (POS)

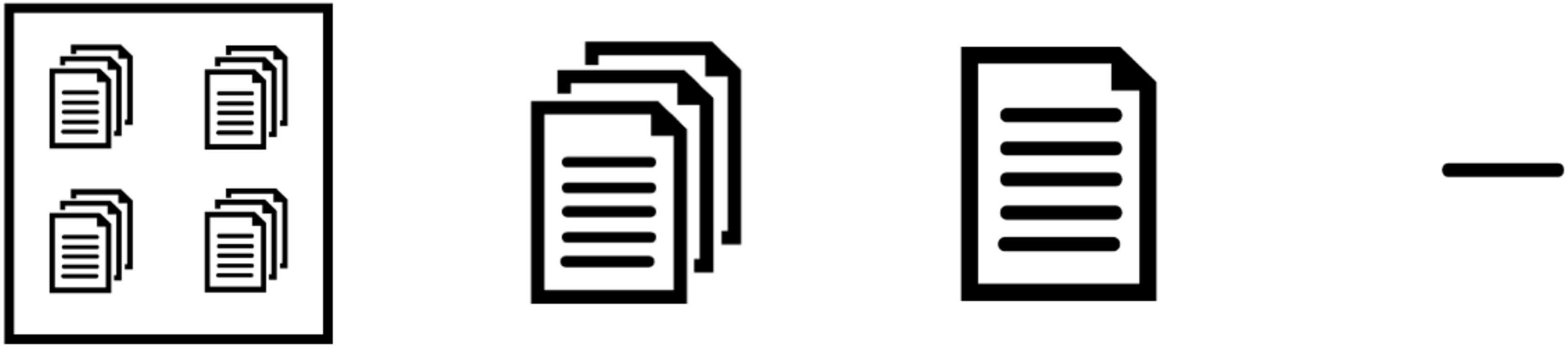
Text Preparation and Pre-processing

- **Uncleaned data** can contain many potential issues, such as misspelled words, incorrect punctuation, improper spacing, etc.,
- using **uncleaned data** can even distort the document's linguistics and inhibit information extraction processes.
- In NLP methods, each word is viewed as a variable (dimension).
- Aim is trying to keep the vocabulary and, thus, the dimensions as small as possible.
- Removing noise from the document can ***reduce*** computational costs and increase NLP models' performance

Language Detection

- **Language detection algorithms** like spacy-langdetect allow the language of documents to be identified.
- It makes sense to perform the language detection at the beginning of the text wrangling process in order to filter just those documents from the entire corpus that correspond to the target language and need to be processed further.

Text Data Hierarchy



Corpora

Corpus

Document

Token

Corpus

- A **Corpus** is defined as a collection of text documents for example a data set containing tweets containing Twitter data is a corpus.
- So **corpus** consists of documents, **documents** consist of paragraphs, **paragraphs** consist of sentences.
- **Sentences** consist of further smaller units which are called **Tokens.**

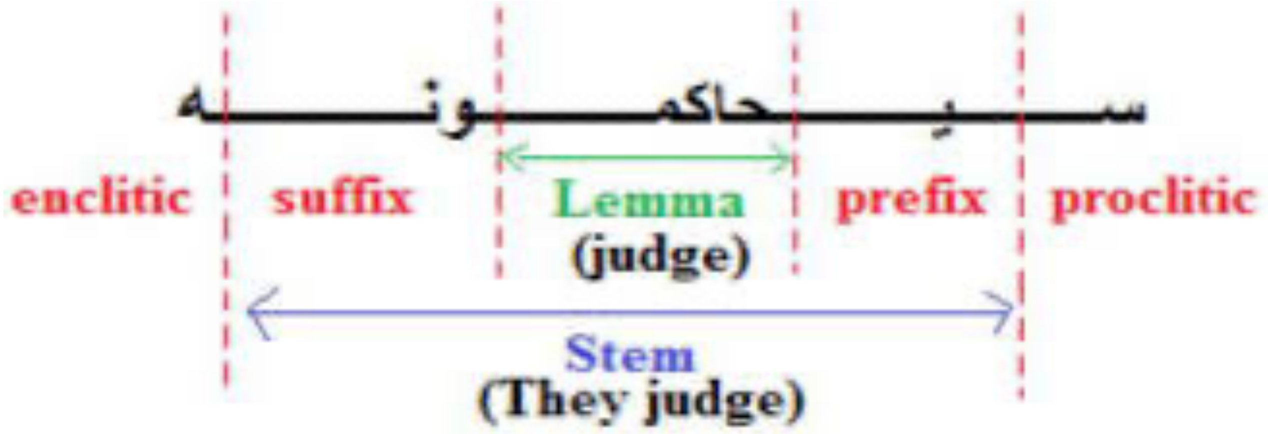
Tokenization

- Languages such as English, German, **Arabic**, or French are space-delimited, meaning that most terms are separated by white spaces.
- For languages like **Thai or Chinese**, however, this is not the case; as these languages do not provide clear boundaries between words, tokenization becomes a challenging task.

Tokenization

- **Tokenization** is a process of splitting a text object into smaller units which are also called tokens.
- The most commonly used tokenization process is **White-space Tokenization :**
  - complete text is split into words by splitting them from white spaces.

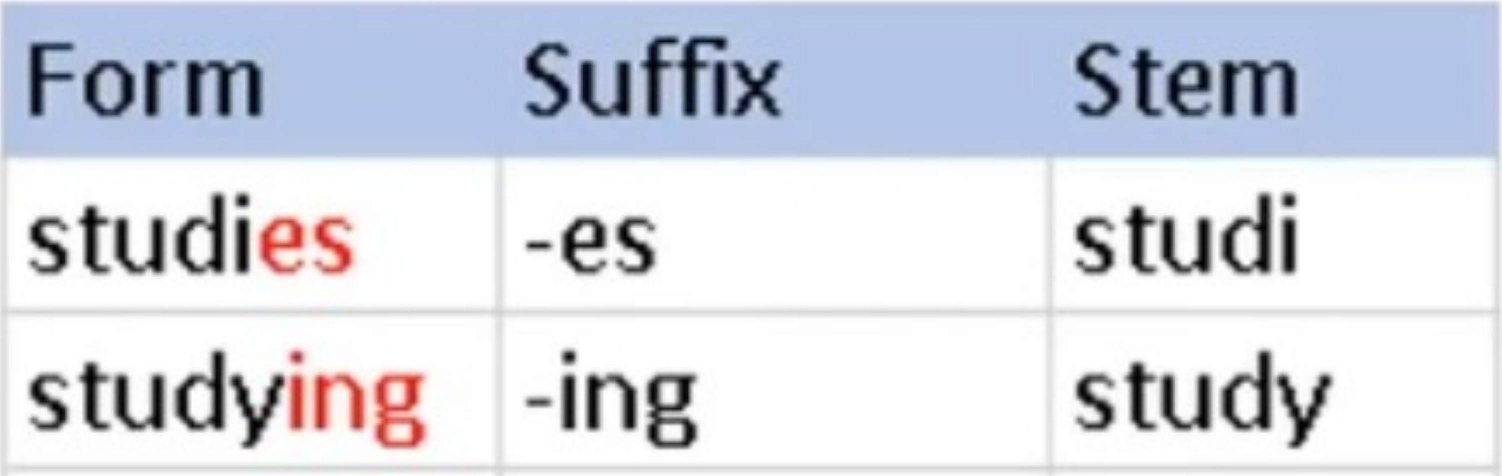
Stemming

- **Stemming** is a rule-based process for removing the end or the beginning of words with the intention of removing affixes. and the outputs will be the stem of the world.
- 

Root = حكم

Meaning = They will judge him

Stemming



Form	Suffix	Stem
studies	-es	studi
studying	-ing	study

Q63. While stemming healed, healing and healer all were reduced to _______________

a. heal

b. healed

c. heale

d. hea

Lemmatization

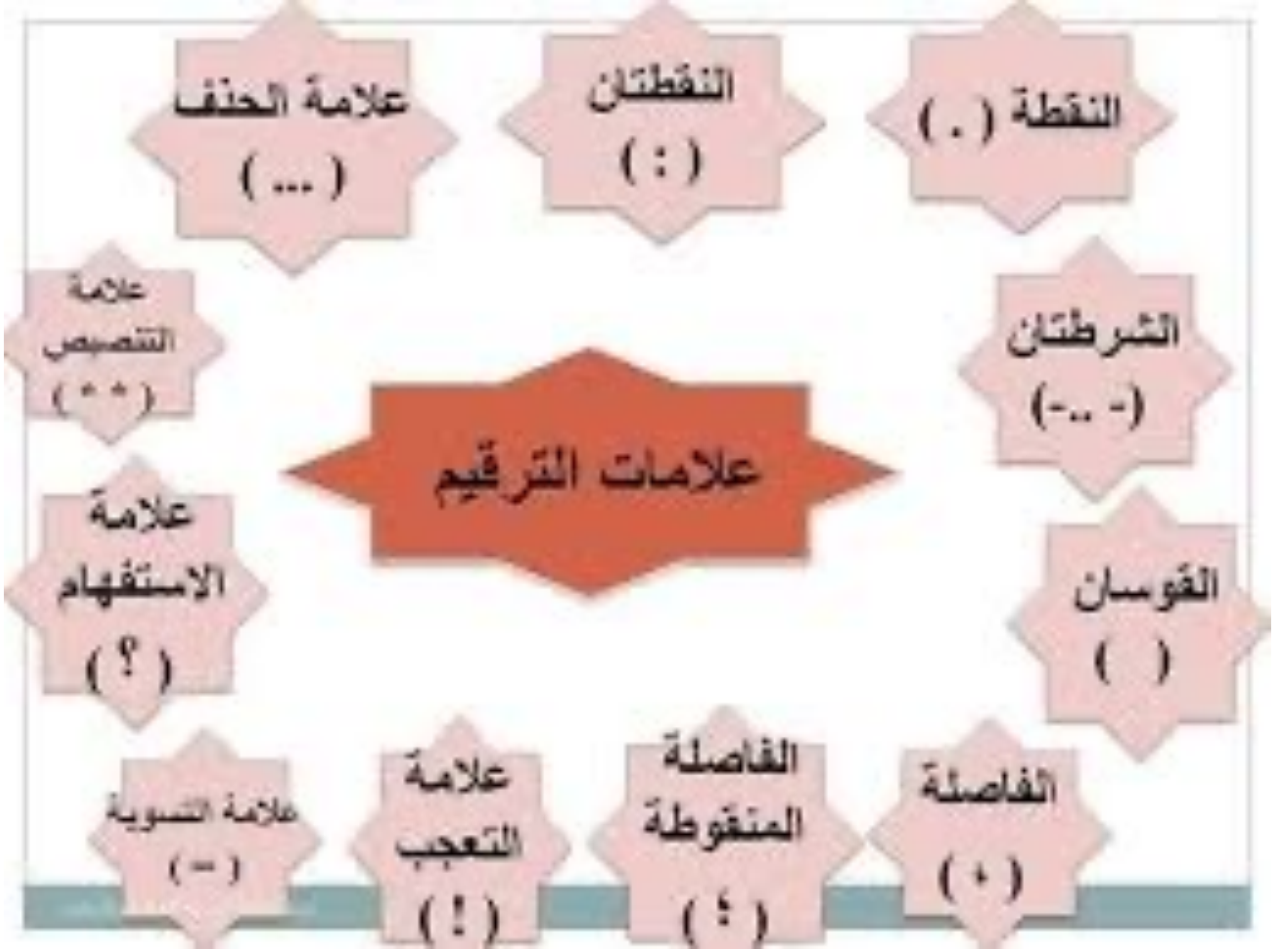
- **lemmatization** returns an actual word of the language, It makes use of vocabulary, word structure, part of speech tags, and grammar relations.
- It is used where it is necessary to get valid words The output of lemmatization is the root word called a lemma.
- For example, the words “running”, “runs” and “ran” are all forms of the word “run”, so “run” is the lemma of all the words.

Lowercasing and Removal of Punctuation

- *Usually*, the transformation of text into lowercase letters and the removal of punctuation are considered the **first pre-processing steps**.
- Lowercase text data is particularly relevant so as to avoid word redundancy.
- For example, when counting words, the terms “Tourism” and “tourism” would be counted as two separate words, leading to an undesired increase in the dimensions of the data.

Lowercasing and Removal of Punctuation

- **Punctuation marks** create noise in the data and do not add value to the analysis, explaining why they should be removed.
- However, in certain situations, it makes sense to analyze individual sentences; in such cases, the corpus should be separated into individual sentences before punctuation marks are removed.



علامات الترقيم

النقطة (.)

النقطتان (:)

علامة الحذف (...)

الشرطتان (- .. -)

القوسان (   )

الفاصلة (،)

الفاصلة المنقوطة (؛)

علامة التعجب (!)

علامة التسوية (=)

علامة الاستفهام (؟)

علامة التنصيص (" ")

Expand Contractions

- **A contraction** is “a shortening of a word, syllable, or word group by omission of a sound or letter”.
- For instance, a contraction like **“we’ll”** is split into two words, **“we” and “will”.**

Removal of Stop Words

- **Stop words** are common words in English such as “a, in, the, can, may”, and so on.
- These words are not useful in NLP analyses because they can be part of any sentence;
- In other words, stop words are not considered keywords in text mining methods.
- Furthermore, they increase vocabulary and hardly provide useful information.

Removal of Stop Words

- There are numerous Python packages that can be used to load stop words, with each package varying in size.
- For example, the **NLTK (Natural Language Toolkit) package** stores a list of stop words for 16 different languages, with 127 stop words specified for English, while the **Stanford NLP package** contains 257 English stop words.
- In most cases, one will use standard stop words but can add individually defined stop words as well.

Removal of URLs, HTML Tags, and Emotions/Emojis

- Regarding **social media posts** from Twitter, Facebook, or Instagram, URLs are often a part of documents.
- However, these limit NLP algorithms from recognizing the actual meaning of a sentence.
- Thus, they are merely noise within a text and should be deleted.
- The same applies to HTML tags as well as emoji's and emoticons.
- Even though they convey information about feelings, we might not want to analyze them and are better off being removed from the data.

Correction of Spelling

- Spelling errors can be seen as another example of increasing the number of features in a vocabulary due to corresponding words being counted twice.
- For example, the words “tourism” and “turism” would be considered two different words.
- To correct spelling errors, many different Python packages using different approaches are available, such as “pyspellchecker”, “SymSpell”, “TextBlob Spell Checker”, etc.
  - However, their execution time and performance differ significantly.

Part of Speech Tagging (POS)

- With **part of speech tagging**, individual tokens are labeled based on their parts of speech.
- This can be achieved using language models that include dictionaries of terms with all their possible parts of speech
- For instance, both **NLTK3 and spaCy** are Python modules that enable extensive POS.

Part of Speech Tagging (POS)

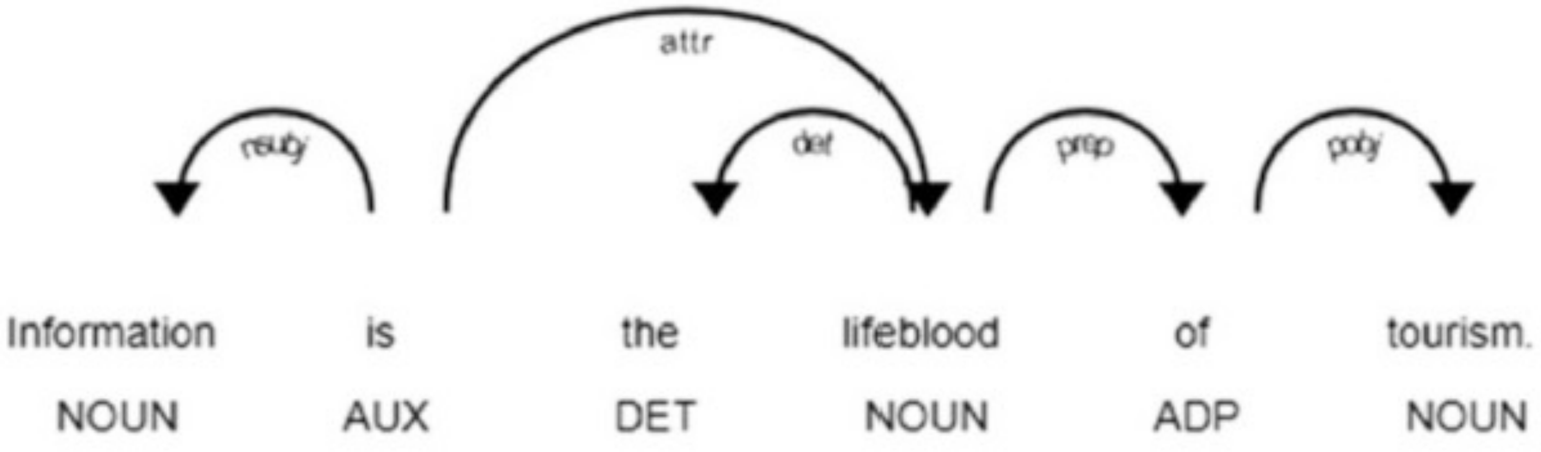


Fig. 5 Part of Speech tagging example

Advantages of Natural Language Processing:

- **Improves human-computer interaction:** NLP enables computers to understand and respond to human languages, which improves the overall user experience and makes it easier for people to interact with computers.
- **Automates repetitive tasks:** NLP techniques can be used to automate repetitive tasks, such as text summarization, sentiment analysis, and language translation, which can save time and increase efficiency.
- **Enables new applications:** NLP enables the development of new applications, such as virtual assistants, chatbots, and question answering systems, that can improve customer service, provide information, and more.

Advantages of Natural Language Processing:

- **Improves decision-making:** NLP techniques can be used to extract insights from large amounts of unstructured data, such as social media posts and customer feedback, which can improve decision-making in various industries.
- **Improves accessibility:** NLP can be used to make technology more accessible, such as by providing text-to-speech and speech-to-text capabilities for people with disabilities.

Disadvantages of Natural Language Processing:

- **Limited understanding of context:** NLP systems have a limited understanding of context, which can lead to misinterpretations or errors in the output.
- **Requires large amounts of data:** NLP systems require large amounts of data to train and improve their performance, which can be expensive and time-consuming to collect.

Disadvantages of Natural Language Processing:

- **Limited ability to understand idioms and sarcasm:** NLP systems have a limited ability to understand idioms, sarcasm, and other forms of figurative language, which can lead to misinterpretations or errors in the output.
- **Limited ability to understand emotions:** NLP systems have a limited ability to understand emotions and tone of voice, which can lead to misinterpretations or errors in the output.

Task

What the different between

- NLP
- NLU
- NLG

Practical (using Python)

- **How to Detect Arabic Language ?**

References

1. Al-Shbiel, A.O., 2017. Arabization and its effect on the Arabic language. *Journal of Language Teaching and Research*, *8*(3), pp.469-475.
2. Al-Salman, A.S., 1996. An Arabic programming environment. *ACM SIGICE Bulletin*, *22*(2), pp.19-25.
3. Shaalan, K., Siddiqui, S., Alkhatib, M. and Monem, A.A., 2018. Challenges in arabic natural language processing. *Computational Linguistics, Speech And Image Processing For Arabic Language*, *4*, p.59.
4. https://morioh.com/p/d596d2d4444d
5. Egger, R. and Gokce, E., 2022. Natural language processing (NLP): An introduction. In Applied Data Science in Tourism (pp. 307-334). Springer, Cham.